Implementing Clustering for Wellness Application

Igors Scapovs

HAAGA-HELIA University of Applied Sciences

2013

HAAGA-HELIA
University of Applied Sciences

Pasila Campus
Business Information Technology

ABSTRACT

October 14, 2013

Authors Igors Scapovs	**Group** BIT
The title of the thesis Implementing Clustering for Wellness Application	**Number of pages and appendices** 21+10
Supervisors Lili Aunimo, Martti Laiho	

Data analysis has always been useful in business, but nowadays with new data being generated every day it has become a must for companies to stay competitive. Data analysis in wellness can be used to gain useful information about people's habits, exercises, nutrition, health, and life quality overall. It can be applied on groups of people or individuals not only to learn about them but also to support them with recommendations and service personalization.

This study implements clustering of questionnaire answers from users of Extensive Life Health-e-Living portal which gives insights about the users of the portal and by looking at the results coach should be able to tell in what condition are portal users and what they need to improve in their lifestyle.

The thesis covers data analysis overall and emphasizes on clustering algorithms and tools which are used in the implementation of the service.

Tools used for the development are Ruby on Rails, on which Health-e-Living portal is based, and Ai4R gem which was used to implement clustering. The results are also visualized in the portal using Rickshaw D3.js library.

Keywords
Analysis, data, clustering, K-means, ruby gems, rails, wellness

Table of contents

1 Introduction

The work completed in this thesis is an implementation of Ruby on Rails application using data analytics tools. The application is processing data taken from questionnaires that are to be filled by future users of Health-e-Living portal. The data is then clustered and visualized and results can be evaluated by caregivers, trainers or other wellness professionals. The implementation is not necessarily specific only to this business. It could be applied for any other business and any data. This implementation can be reused or reconfigured for any other purpose just with a little knowledge of coding.

“Health-e-Living is a Personal Health Coach Portal for all citizens that induces better health habits for those under risk and replaces bad habits for those suffering already of chronic illness.” (Extensive Life)

Extensive Life Health-e-Living portal is an online service to support different lifestyles, demographics, and health conditions of individuals and communities.

As a part of the service Health-e-Living it will help:

- Understanding all factors of your health spectrum.
- Defining your health values and goals.
- Achieving goals through a personalized lifestyle program.
- Improving health by inducing behaviors towards healthier habits.
- Keeping you in track of a balanced and healthy lifestyle.

(Extensive Life)

“The combination of increasingly complex world, the vast proliferation of data, and the pressing need to stay one step ahead of the competition has sharpened focus on using analytics within organizations.” (Steve LaValle, Michael Hopkins, Eric Lesser, Rebecca Shockley and Nina Kruschwitzi 2010, 1)

The use of data analytics is a differentiator in business nowadays; it can help guide future strategies, give insights, and help make decisions. (Steve LaValle, Michael Hopkins, Eric Lesser, Rebecca Shockley and Nina Kruschwitzi 2010, 2)

1.1 Background for the study

Extensive Life had a vision and a plan that they will need to use data analytics to support, improve, and revolutionize their services on the Health-e-Living platform.

1.2 Purpose of the study

This study would give Extensive Life a starting point for implementing different services with data analytics with the purpose to automatize services and to gain valuable information from the data.

Another purpose is to apply data analysis to encourage people to make lifestyle choices that would make them remain healthy and prevent developing diseases, which is also the main purpose of Health-e-Living portal.

This study will also help people to re-implement this Ruby on Rails application using the same techniques applied in this work and use it for their own needs.

This work can also be used as an educational material about data analysis.

1.3 Research Problems

- How to cluster the portal users in different groups to personalize the service?
- How can data analysis support business in wellness industry?

2 Data Analysis

This part covers data analysis techniques, like clustering, classification, association rules and emphasizes on clustering and clustering algorithms. There are many algorithms available; this part covers only the most common ones giving short description and some examples to provide understanding, and practical application of data analytics for people with no prior knowledge in this field. Data analysis can be challenging to learn and apply, but it can unveil interesting and valuable information.

Data analysis can be applied to reveal the relationships, patterns, trends in the data. It can also reveal the level of trust for the results. It means that results can be compared to other results, for example, from control group or from comparison group, and then conclusions can be drawn. The point is to get an accurate assessment and results that are useful. (Phil Rabinowitz & Stephen Fawcett)

The amount of data generated has risen steadily every year and more data is stored in digital formats. The challenge is how to deal with all the data and which data can provide value to business. Data itself may not bring value but with help of analytics it can uncover valuable business information. (Oracle Corporation 2013)

2.1 Data Analysis in Wellness

> "There is a growing interest in changing the way individuals prevent illness and treat diseases. Improved medical practices, holistic approaches, and new intellectual and spiritual philosophies have contributed to the movement towards an integrated approach to healthcare." (Metro Denver 2012, 1)

There were around 289 million active wellness consumers in the world's top 30 industrialized nations alone as of year 2010. Wellness is multi-dimensional and holistic, integrating physical, mental, spiritual, and social approaches. Wellness is also complementary and proactive and is focusing more on preventing sickness and improving quality of life rather than treating an existing illness. It's also consumer driven and they can make choices to their preference. (PRWeb 2010)

Main components of wellness can be seen in Figure 1. These components can be saved as digital data and then analyzed. Valuable information can be found when these components are correlated, for example, after a period of time we could tell how sleep affects everyday activity, workouts, nutrition, or how nutrition affects sleep and body measurements.

Figure 1: Dimensions of Wellness (Health Graph Blog 2013)

According to Ki Mae Heussner, collection of data has led to new insights on how diet and exercise can improve health, and those insights can help generate customized health programs. It has already helped many patients to improve their heart health and lose weight. Combining data from genetic tests with metabolic assessments and data collected from internet and mobile devices has helped to achieve that. Corporate wellness is also gaining popularity and importance. Companies start using digital services for wellness to maintain employee health and to boost their productivity. (Ki Mae Heussner 2013)

In one study K-means algorithm was used to identify natural groupings of individuals who choose to eat similar foods. In combination with other data it was identified that demographic, behavioral, and health factors were associated with the choice of foods. Individuals who were eating the healthiest were reported to have physical illness less likely than

those with poor food choices. (Justin B. Dickerson, Matthew Lee Smith, Rhonda Rahn, Marcia G.Ory)

Data analysis identified that increased consumption of sugared beverages is associated with overweight in adults and children. Sugared beverages were also shown to negatively affect nutritional consumption. Fast food was also shown to cause overweight and obesity. A diet low in fruit and vegetable is associated with higher risks of ischemic stroke, lung cancer, stomach cancer, colorectal cancer and other types of cancer. In opposite higher consumption of fruits and vegetables reduced the risk of cancer and also reduced the risk of hypertension. (Justin B. Dickerson, Matthew Lee Smith, Rhonda Rahn, Marcia G.Ory)

Many adults and children still consume diets low in fruits and vegetables and with help of data analysis it is easier to find factors that affect health and also promote people to making healthier food choices. (Justin B. Dickerson, Matthew Lee Smith, Rhonda Rahn, Marcia G.Ory)

2.2 Classification

Classification is used to predict target class for each item in the data. The target values are already known and this data is used to train the machine. To test how precise are predictions, part of the data is used to build the model and part of the data is used for comparing predicted values against known values.

For example, it could be used to identify loan applications as low, medium, or high credit risks. The classification rules for predicting to which class a loan applicant belongs could be developed from history data of other loan applicants. The algorithm would predict to which class customer belongs with probability. Classification can be used for customer segmentation, business modelling, marketing, credit analysis, and biomedical and drug response modelling. (Oracle Corporation 2008, 5-1.)

case ID | predictors | target

CUST_ID	CUST_GENDER	EDUCATION	OCCUPATION	AGE	AFFINITY_CARD
101501	F	Masters	Prof.	41	0
101502	M	Bach.	Sales	27	0
101503	F	HS-grad	Cleric.	20	0
101504	M	Bach.	Exec.	45	1
101505	M	Masters	Sales	34	1
101506	M	HS-grad	Other	38	0
101507	M	< Bach.	Sales	28	0
101508	M	HS-grad	Sales	19	0
101509	M	Bach.	Other	52	0
101510	M	Bach.	Sales	27	1

Figure 2: Sample Build Data for Classification (Oracle Corporation 2013, 5-2)

In the Figure 2 historical data of customers is shown. AFINITY_CARD column values can be 0 or 1, with 0 meaning that customer did not increase spending with affinity card and 1 meaning that customer increased spending with affinity card. After training, machine will be able to predict if new a customer will increase spending with affinity card or not. For example, last 2 rows could be excluded from training, then the target values of those 2 rows are predicted and then they are compared to the actual values.
(Oracle Corporation 2008, 5-2.)

2.3 Association rules

Association rules are *if* and *then* statements that help to uncover relationships between seemingly unrelated data. *Support* and *confidence* are used for criteria. *Support* indicates how frequently items appear and *confidence* indicates the number of times if/then statements have been found to be true. (Margaret Rouse 2011)

Association rules are often used to predict what customers are going to buy in the supermarket. It is called *market-basket analysis*. For example, if a customer buys cereals then it is very likely that he will buy milk as well. By knowing this, store can be organized and cross-sell can be achieved. Cross-sell is a marketing method to sell complementary products to a product that customer has already purchased. Customer buys an iPod, but in addition he buys Apple earphones; that is an example of cross-sell, same idea as in the previous example with cereals and milk. Association rules can be useful in marketing, sales promotions, for discovering business trends, also in e-commerce for Web page per-

sonalization. An example of Web page personalization is that user who visits pages A and B is 70% likely to visit also page C, and based on this rule, a dynamic link could be placed for users who are likely to be interested in page C.
(Oracle Corporation 2008, 8-1; Bustos 2009.)

The difference between classification rules and association rules is that association rules can predict more than just a class. (Witten, Frank & Hall 2011, 12.)

“Association rules differ from classification rules in two ways: They can “predict” any attribute, not just the class, and they can predict more than one attribute’s value at a time. Because of this there are far more association rules than classification rules, and the challenge is to avoid being swamped by them.” (Witten, Frank & Hall 2011, 41.)

Iris dataset usually contains numeric values of petal and sepal height and length in centimeters. Association rules usually involve only nonnumeric attributes; thus, you wouldn’t normally look for association rules in the iris dataset. (Witten, Frank & Hall 2011, 41.)

2.4 Clustering

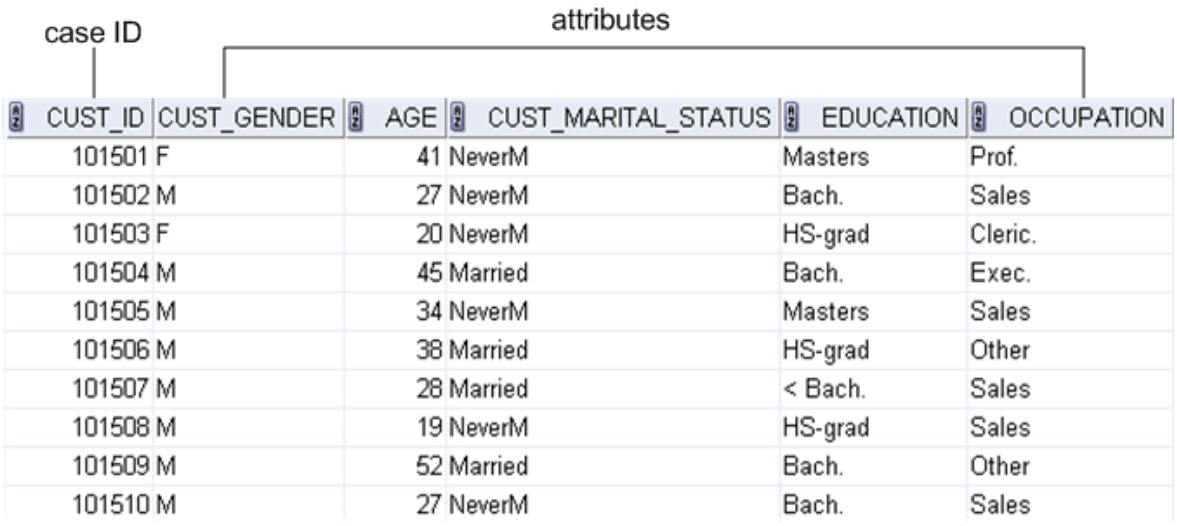
Clustering is used to organize items into groups (clusters) where similar items belong to the same cluster. There can be 1 or many clusters and using this method useful information can be found. The similarity between items in the group can be defined by a function or a formula. (Prabu Arumugam 2010b)

Clustering can be used in:

- marketing to find customers with similar behaviour
- biology to classify biological species by their features
- libraries for book ordering
- in earthquake studies to identify dangerous areas

(Matteo Matteucci)

Figure 3 below shows an example model of data for clustering. Compared to classification there is no target column. The goal is to organize these items into natural groups. (Oracle Corporation 2008, 7-4.)



CUST_ID	CUST_GENDER	AGE	CUST_MARITAL_STATUS	EDUCATION	OCCUPATION
101501	F	41	NeverM	Masters	Prof.
101502	M	27	NeverM	Bach.	Sales
101503	F	20	NeverM	HS-grad	Cleric.
101504	M	45	Married	Bach.	Exec.
101505	M	34	NeverM	Masters	Sales
101506	M	38	Married	HS-grad	Other
101507	M	28	Married	< Bach.	Sales
101508	M	19	NeverM	HS-grad	Sales
101509	M	52	Married	Bach.	Other
101510	M	27	NeverM	Bach.	Sales

Figure 3: Sample Build Data for Anomaly Detection (Oracle Corporation 2013, 6-3)

Clustering and classification are very common techniques used in data analytics. Although, they look very similar at first as they both organize items into groups, but they are actually different. Classification is supervised, so classes may be already known and rules have to be defined by which the classifier algorithm will classify items. Clustering is unsupervised, so the machine will decide into which groups items are assigned. (Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze 2009, 349-350)

Hierarchical clustering is a technique that produces hierarchical structure of clusters. There is a top level cluster that divides into smaller clusters and then the smaller clusters divide into even smaller clusters. This type of clustering is often used to cluster biological species. (Ian H. Witten, page 81)

Figure 4 shows an example what Hierarchical clustering is. The diagram is called dendrogram. Clustering dendrogram on the left is exactly the same as the one on the right; it is just represented in a different way. One big cluster can be seen, and within, more clusters that break down into smaller clusters.

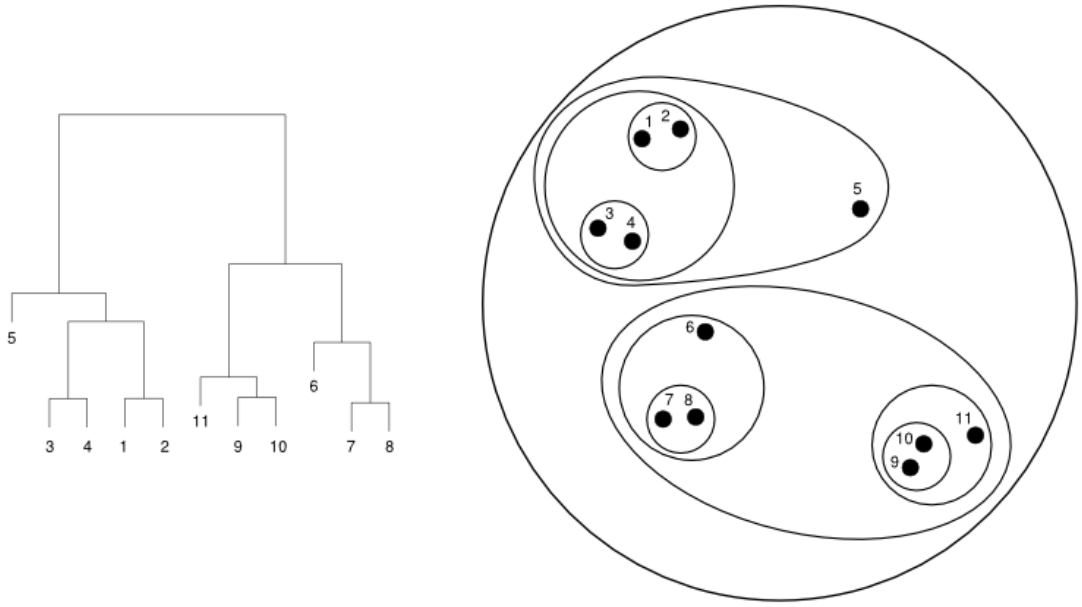


Figure 4: Dendrogram (Răzvan Musăloiu-E 2009)

> ”In a agglomerative procedure, each object starts out as its own cluster. In the subsequent steps, the two ‘closest’ clusters/objects are combined into a new aggregate cluster, thus reducing the number of clusters by one in each step. Eventually, all objects are grouped into one large cluster.” (Western Michigan University, IV.B-6)

> ”In a divisive procedure, the process proceeds in the opposite direction: we start out with one large cluster containing all objects; in the subsequent steps, the objects that are most dissimilar are split off and turned into smaller clusters; this process continues until each object forms a cluster of itself.” (Western Michigan University, IV.B-6)

“Many variants of the basic k-means procedure have been developed. Some produce a hierarchical clustering by applying the algorithm with k = 2 to the overall dataset and then repeating, recursively, within each cluster.” (Witten, Frank & Hall 2011, 141.)

2.4.1 K-Means

K-means is a clustering technique that requires that the number of clusters is specified. K is a parameter which represents the number of clusters. K points are chosen randomly by the algorithm as cluster centers according to Euclidiean distance metric, then the centroid of each cluster is calculated. Centroid is a central point in the cluster that decides which

items belong to that cluster. The process can be repeated several times until most stable centroid is found. (Witten, Frank & Hall 2011, 139.)

“Suppose you are using k-means but do not know the number of clusters in advance. One solution is to try out different possibilities and see which is best. A simple strategy is to start from a given minimum, perhaps k = 1, and work up to a small fixed maximum. Note that on the training data the “best” clustering according to the total squared distance criterion will always be to choose as many clusters as there are data points!” (Witten, Frank & Hall 2011, 274.)

2.4.2 Divisive Clustering Procedure - DiAna

A divisive clustering procedure, named DiAna (short for Divisive Analysis Clustering), was developed by Kaufman, L. and Rousseeuw, P.J. (Finding Groups in Data: An Introduction to Cluster Analysis, 1990, Wiley, New York.)

> ”The algorithm is as follows:
>
> - It starts with a single cluster, the entire set of n objects.
> - In each step, the cluster with largest diameter is selected and is to be divided (splintered) into two clusters. Here the diameter of a cluster is defined as the maximum distance or dissimilarity (i.e., minimum similarity) among all objects within the cluster. An object within this cluster having largest average dissimilarity to other objects within the cluster is identified. This object initiates ‘*splinter group*.’ An object within this cluster is reassigned to the splinter group if it is closer to the splinter group than to the ‘old party.’ Consequently, at the end of the step, the cluster is divided into two new clusters.
> - The above step is repeated until n clusters are formed.”
>
> (J. C. Wang 2009, IV.B-12)

2.4.3 Ward's Method

"This is an alternative approach for performing cluster analysis. Basically, it looks at cluster analysis as an analysis of variance problem, instead of using distance metrics or measures of association.

This method involves an agglomerative clustering algorithm. It will start out at the leaves and work its way to the trunk, so to speak. It looks for groups of leaves that it forms into branches, the branches into limbs and eventually into the trunk. Ward's method starts out with n clusters of size 1 and continues until all the observations are included into one cluster.

This method is most appropriate for quantitative variables, and not binary variables.
Based on the notion that clusters of multivariate observations should be approximately elliptical in shape, we assume that the data in each of the clusters will be realized in a multivariate distribution. Therefore, it would follow that they would fall into an elliptical shape when plotted in ap-dimensional scatter plot." (Pennsylvania State University 2004)

3 Case Implementation

Data analytics for Extensive Life Health-e-Living Web Portal must support wellness coaching service and provide reports about the users of the portal, including their health, nutrition, physical activity, mood, stress, sleep, and other indicators of health status.

The objective of this implementation is to provide coaches with insights about people's health and behaviours to support coaches in their decisions/actions on how to improve individual person's health.

The picture (Figure 5) shows service flow of Health-e-Living:

Figure 5: Health-e-Living service flow

The service flow of Health-e-Living includes:

- Profiling of users when they register in the portal
- Planning diet, exercise, sleep and other parts of healthy lifestyle
- Monitoring of exercise, diet, sleep and other
- Find causes for failing on plan

Data analysis can support each of these services. In profiling users can be clustered by their answers to questionnaire, their location, age. In planning users can get recommendations for their exercise program or suggest healthy diets. In monitoring analytics can do predictions when results can be achieved or what has to be changed in lifestyle regime. Finding causes for failing on plan can help find correlations in diet, exercise, sleep and other components, and tell the user what are the causes for bad sleep, for example. These are just some examples how data analytics can be applied on Health-e-Living services.

The implementation emphasizes on the clustering part. When users first register in the portal they are asked to fill out a questionnaire form about their health, diet, and habits. After they have submitted the form, clustering of users can be performed. This allows to see how users of the portal are distributed and to learn if there are more users who need to improve their diet, adjust exercise or change habits. The implementation can also be adjusted to cluster and visualize any other data and can be used in any business field. The implementation focuses only on open-source software.

3.1 Tools and setup

Most part of the Extensive Life Health-e-Living Web Portal is developed with Ruby on Rails. For this reason implementation of clustering was developed using Ruby on Rails and using Ruby gems was the most convenient way to accomplish that. Same or similar techniques could be applied using other web development frameworks, but that would require research on tools for that specific framework. There are other ways to do analytics other than using a Ruby gem, for example, using another software product like Knime and integrating it with the application or using a bridge (RsRuby gem for Ruby) to do analytics with R. There are advantages and disadvantages for any approach. Using an analytics gem that has to be added to the gemlist file in the Ruby on Rails application was the most convenient way for Health-e-Living Web Portal. It is easy, straightforward, has enough functionality in order to do data analysis, and it also does not require to install extra software on a production server (except for the gem). Knime and R are also great tools and choosing which to use depends only on the skills and freedom of a developer. Knime and R were the candidates for this implementation and were found to be better than most of the open-source software that is available on the internet. More detailed comparison between Knime, R, and Ruby gems is provided below.

Knime

Website: http://www.knime.org/

“KNIME, pronounced [naim], is a modern data analytics platform that allows you to perform sophisticated statistics and data mining on your data to analyze trends and predict potential results. Its visual workbench combines data access, data transformation, initial investigation, powerful predictive analytics and visualization.” (KNIME.com AG 2013)

Advantages:

- Lots of functionality (pre-processing, modelling, different algorithms, visualization etc.)
- Extensible (integrates with R, Weka, Groovy, Python, Matlab, and other software)
- Graphical interface

Disadvantages:

- Complex and difficult to learn
- Difficult to implement and integrate with web applications
- Has to be installed on a production server

Using bridge to R.

Website for R: http://www.r-project.org/
Website for RsRuby: (http://rubygems.org/gems/rsruby)

> ”R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.” (R Project)

Ruby gem RsRuby can be used as a bridge to access R. It is an interpreter that allows R methods to be called from Ruby and data passed between Ruby script and the R interpreter. (Alex Gutteridge, Ben J Woodcroft 2009)

Advantages:

- Almost unlimited possibilities for data analysis

Disadvantages:

- Requires to learn R and requires a lot of manual work
- No graphical interface
- R has to be installed on a production server

Ruby gems

Advantages:

- There is no need to install extra software on the production server (just gem)
- Takes less space

- Easy integration
- Easy to use

Disadvantages:

- Limited functionality (there can be a gem that implements only 1 algorithm or it can't read data straight from the database)
- No graphical interface

Taking into consideration that this was a starting point for using data analytics for Health-e-Living Web Portal, in Table 1 some of the Ruby data analysis gems are described and during this research Ai4r was found to be the best, it has lots of functionality, updated recently, and well documented, including good practical examples.

Gem	Description	Documentation	Last Update	Website
clusterer	Gem for clustering text data with various algorithms	Well documented with examples.	March 22, 2007	http://rubyforge.org/projects/clusterer/ https://www.ruby-toolbox.com/projects/clusterer
classifier	Gem for classiffying text using Bayes algorithm. Easy to use.	Well documented with examples.	2010	https://www.ruby-toolbox.com/projects/classifier
stuff_classifier	Naive Bayes based text classifier. Easy to use. Ignore list for unwanted words.	Well documented with examples.	2013	https://github.com/alexandru/stuff-classifier/ https://www.ruby-toolbox.com/projects/stuff-classifier
kmeans	Gem implementing K-means algorithm. Can be used to cluster data with multiple columns.	Poorly documented, bu some examples can be found	2013	https://github.com/id774/kmeans https://www.ruby-toolbox.com/projects/kmeans

Gem	Description	Documentation	Last Update	Website
		inside the gem's 'spec' folder.		
Ai4r	Implementing many different algorithms.	Very well documented, including – installation, examples, algorithm descriptions etc.	2013	http://www.ai4r.org/ https://www.ruby-toolbox.com/projects/ai4r
RsRuby	RsRuby is a bridge to R that allows to write code in Ruby script. It's limited only with the functionality of R.	Well documented, but missing good examples.	February 2009	http://rubygems.org/gems/rsruby
RinRuby	Similar to RsRuby, doesn't need R compilation, but is considered slower than RsRuby.	Well documented with good examples.	July 2012	https://sites.google.com/a/ddahl.org/rinruby-users/
RsRuby	RsRuby is a bridge to R that allows to write code in Ruby script. It's limited only with the functionality of R.	Well documented, but missing good examples.	February 2009	http://rubygems.org/gems/rsruby

Table 1. Popular data analytics gems from www.ruby-toolbox.com.

Ai4r Gem

Ai4R was chosen from the gems because it had examples, was fairly easy to use and had all the functionality needed for the implementation of clustering with K-Means, Diana or Ward's method. It also has implemented other algorithms (classification, neural networks, genetic algorithms etc.) that may become of use in future implementations.

3.2 Pre-processing data

To successfully cluster and visualize data it requires knowing specifically which tools to use, the data, data input/output formats and the whole process from beginning to the end. The data must be pre-processed first in order to be passed to the clustering function. The data from questionnaire answers is stored like seen in the picture in Appendix 1. *value* column in the table corresponds to the answer option submitted by the user.

The task is to convert these values into X's and Y's. Methods are created to convert these values from the table to a number specified by the developer. In this case there are 3 dictionaries - calculate_good_diet, calculate_bad_diet and calculate_exercise. Methods can be seen in Appendix 3. These methods convert *value*'s from database to numbers specified by the developer, these numbers can be also called points and they will determine X and Y values.

X and Y values are calculated by a function calculate_points seen in Appendix 3. This function sums the values contained within the array. This function is called inside of the function which has a dictionary, for example, calculate_good_diet and calculate_exercise, and utilizing the array of converted answers. As a result X and Y is generated, a single array is created, like [25,15]. This array will determine where this data point is positioned in the graph.

Method preprocess_assessments takes 2 functions as parameters, for example, calculate_good_diet and calculate_exercise then it uses them to pre-process data from *Answers* table and it returns an array containing all data points generated from the data. This array is ready to be passed to the clustering algorithm.

3.3 Clustering

Clustering is achieved with K-Means algorithm which was provided by Ai4R gem, it also has other algorithms including DiAna and Ward's method and they also can be used. K-Means was used because it is considered a popular and generally useful/precise algorithm, also compared to Diana it was performing hundreds of times faster and producing results similar to those with DiAna. Performance was out of scope of this thesis, so was the precision of results as real data was not available for testing. Clustering of 10000 data points was done under a second with K-Means, whereas for DiAna it took minutes, which is a rough estimation of time it took when run on a home laptop computer. If necessary the code can be easily changed to utilize any other algorithm provided by Ai4r gem, just by changing one line in the code:

```
clusterer = Ai4r::Clusterers::KMeans.new.build(data_set, 3)
```

KMeans here can be changed to DiAna. The number of clusters is also defined here, and 3 clusters were used in this implementation because it was agreed with Extensive Life that users will be grouped into 3 categories and at the moment it is enough. Later on it could be parameterized so that number of clusters could be defined in the view by the user. The same line of code actually does the clustering and saves data to the *clusterer* variable.

Clustering data is actually the easiest part of the whole implementation. The most difficult and time consuming parts is transforming and retrieving data, talking about the technical side. From the business side the most difficult would be defining requirements for the application.

The code for clustering and some of data transformations can be seen in Appendix 4, notice that controller is using helper methods from Appendix 3 which also includes data transformations.

After clustering is complete and *clusterer* variable contains data, it is transformed with format_for_graph method. This method sorts the data and does mapping, then pushes data to the array. This array then can be passed to the graph for visualization.

3.4 Visualizing

Researching on different visualization frameworks drove to a conclusion that using D3.js JavaScript library was the best option. It is open-source, has many tutorials, contributors, examples, and overall it is a very powerful tool to visualize data. The data is displayed in an SVG tag in the browser, it is very scalable, fast, and compared to PNG or JPEG file formats is interactive. Also it is a lot faster compared to Flash. Even though SVG file format is not supported in older browsers.

> "D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation." (Michael Bostock 2012)

For this implementation Rickshaw D3.js plugin was used. It can be downloaded from http://code.shutterstock.com/rickshaw/. This plugin was chosen among many others because it looked great, had possibility to change data point colors, to hover on data points, and had lots of flexibility to change/customize the graph for the needs. The code implementation of visualization can be seen in Appendix 5. When user clicks *Run* button the process of pre-processing, clustering, data transformations starts and data is passed from the controller to the view. In the view data points are colored according to the cluster they belong.

Clustering and visualization was actually done before pre-processing part with test data that was created with SQL script which can be found in Appendix 6. Script filled a table just with X and Y values which were easier to pass for clustering and any amount of data could be created. It could have been done also with the real *Answers* table, but the script would be too complex and time consuming. The graph was tested with 10000 data points had no problem with performance. Loading time is almost instant.

4 Summary

The graph seen in Figure 6 could be called as the end result of this work. It shows 300 data points clustered with K-Means. When a new user registers and fills out the form another data point will appear in this graph. Trainers and users who have access to this page are able to see clusters of users according to the questionnaire answers.

The graph is supposed to add a data point for each questionnaire that a user fills out (in this case 1 when the user registers for the first time). Data points are clustered with K-Means algorithm and each clustered appears in different color. 3 Clusters were used in this implementation.

Results from questionnaire can be useful to learn about the portal users overall according to the initial assessment and, for example, see how many users have a bad diet or how many users need to exercise more.

This implementation probably will be developed further. Next steps would be to make it more parameterized, meaning that a user can choose more options for reporting. Finding a specific user in the graph and also displaying a table showing amount of male/female and age groups contained within each cluster. It could also be developed to retrieve types of information other than questionnaire data.

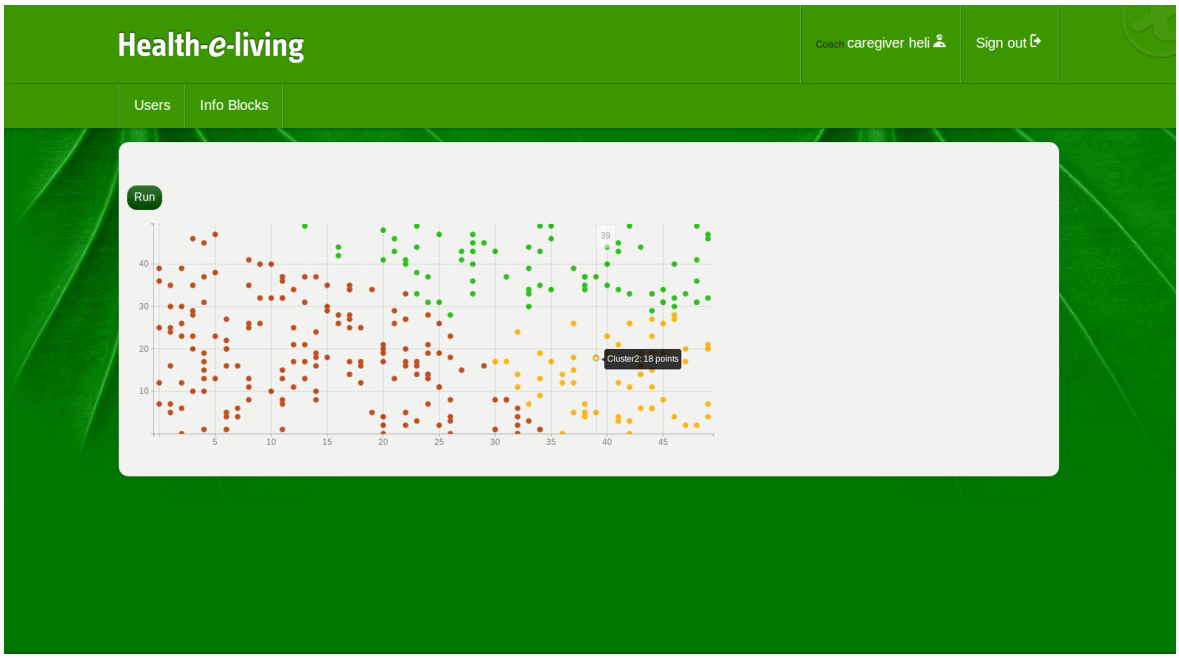


Figure 6: Graph of 300 records clustered with K-Means

4.1 Further development

Data analysis has a wide range of use and this implementation was only a small part of what can be achieved with data analysis. The implementation clustered only data from questionnaires, but it could be customized or further developed to cluster any other data and be used for creating automatic personalized recommendations based on the results of clustering.

Extensive Life will also implement other types of services in the future using data analytics, as there are so many possibilities and value of using data analytics.

4.2 Evaluation

The work is evaluated by Extensive Life as well as the importance of this work on their further implementations.

Bibliography

Alex Gutteridge, Ben J Woodcroft. 2009. URL: http://rubygems.org/gems/rsruby. RSRuby. Quoted: 9.10.2013.

Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze. 1st April 2009. Cambridge University Press. URL: http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf. Quoted: 09.10.2013.

Extensive Life. 2013. Heatlh-e-Living. URL: http://health-e-living.com/. Quoted: 13.10.2013.

Health Graph Blog.2013. URL: http://bloghealthgraph.files.wordpress.com/2012/06/health.png. Last accessed: 7.10.2013.

Ian H. Witten, Eibe Frank, Mark A. Hall. 2011. Data Mining, Practical Machine Learning Tools and Techniques. Third Edition. Elsevier. USA.

Justin B. Dickerson, Matthew Lee Smith, Rhonda Rahn, Marcia G.Ory. Behavioral Cluster Analysis of Food Consumption: Associations with Comparatively Healthier Food. URL: http://archive.ispub.com/journal/the-internet-journal-of-nutrition-and-wellness/volume-11-number-1/behavioral-cluster-analysis-of-food-consumption-associations-with-comparatively-healthier-food-choices.html#sthash.6lxhDw8N.dpuf. Quoted: 7.10.2013.

J. C. Wang. 2012. Western Michigan University. URL: http://www.stat.wmich.edu/wang/561/classnotes/Grouping/Cluster.pdf. Quoted: 13.10.2013.

Ki Mae Heussner. 26th June 2013. Data-driven health practice MDRevolution launches web-based employee wellness service. URL: http://gigaom.com/2013/06/26/data-driven-health-practice-mdrevolution-launches-web-based-employee-wellness-service/. Quoted: 7.10.2013.

KNIME.com AG. 2013. KNIME Desktop. URL: http://www.knime.org/knime. Quoted: 9.10.2013.

Linda Bustos. 6th July 2009. Cross-Sells and Upsells: What is the Diff? URL: http://www.getelastic.com/defining-cross-sell-upsell/. Quoted: 11.6.2013.

Margaret Rouse. February 2011. Association rules (in data mining). URL: http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining. Quoted: 11.6.2013.

Matteo Matteucci. Politecnico di Milano. A Tutorial on Clustering Algorithms. URL: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/. Quoted: 13.10.2013.

Metro Denver. January 2012. HEALTHCARE AND WELLNESS.Metro Denver and Northern Colorado Industry Cluster Profile. URL: http://digitalmavrik.com/prime/wp-content/uploads/2012/08/Metro-Denver-EDC_HealthcareWellness_Industry-Profile.pdf. Quoted: 7.10.2013.

Michael Bostock. 2012. URL: http://d3js.org/. Data-Driven Documents. Quoted: 9.10.2013.

Oracle Corporation. May 2008. Oracle Data Mining Concepts, 11g Release 1(11.1). URL: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129.pdf. Quoted: 11.6.2013.

Oracle Corporation. March 2013. Big Data Analytics. URL: http://www.oracle.com/technetwork/database/options/advanced-analytics/bigdataanalyticswpoaa-1930891.pdf . Quoted: 7.10.2013.

Pennsylvania State University. 2004. Ward's Method. URL: http://sites.stat.psu.edu/~ajw13/stat505/fa06/19_cluster/09_cluster_wards.html. Quoted: 13.10.2013.

Phil Rabinowitz and Stephen Fawcett. Collecting and Analyzing Data. URL:http://ctb.ku.edu/en/tablecontents/chapter37/section5.aspx. Quoted: 7.10.2013.

Prabu Arumugam. 3rd February 2010. K-Means Algorithm. URL: http://www.codeding.com/?article=14. Quoted: 11.6.2013.

PRWeb. 26th May 2010. Wellness No Passing Fad: Global Market Estimated at Nearly $2 Trillion, According to Landmark Study Unveiled at Global Spa Summit. URL: http://www.prweb.com/releases/GlobalSpaSummit/SRI/prweb4052824.htm. Quoted: 7.10.2013.

Răzvan Musăloiu-E. 2009. Dendrogram. URL: http://cs.jhu.edu/~razvanm/fs-expedition/hclust-example.png. Last accessed: 7.10.2013.

Steve LaValle, Michael Hopkins, Eric Lesser, Rebecca Shockley and Nina Kruschwitzi. 2010. IBM. Analytics: The new path to value. URL: http://www-935.ibm.com/services/uk/gbs/pdf/Analytics_The_new_path_to_value.pdf. Quoted: 13.10.2013.

The R Project for Statistical Computing. URL: http://www.r-project.org/. Quoted: 9.10.2013.

Western Michigan University.2012. Clustering Analysis. URL: http://www.stat.wmich.edu/wang/561/classnotes/Grouping/Cluster.pdf. Quoted: 9.10.2013.

Appendices

Appendix 1: Model

Model:

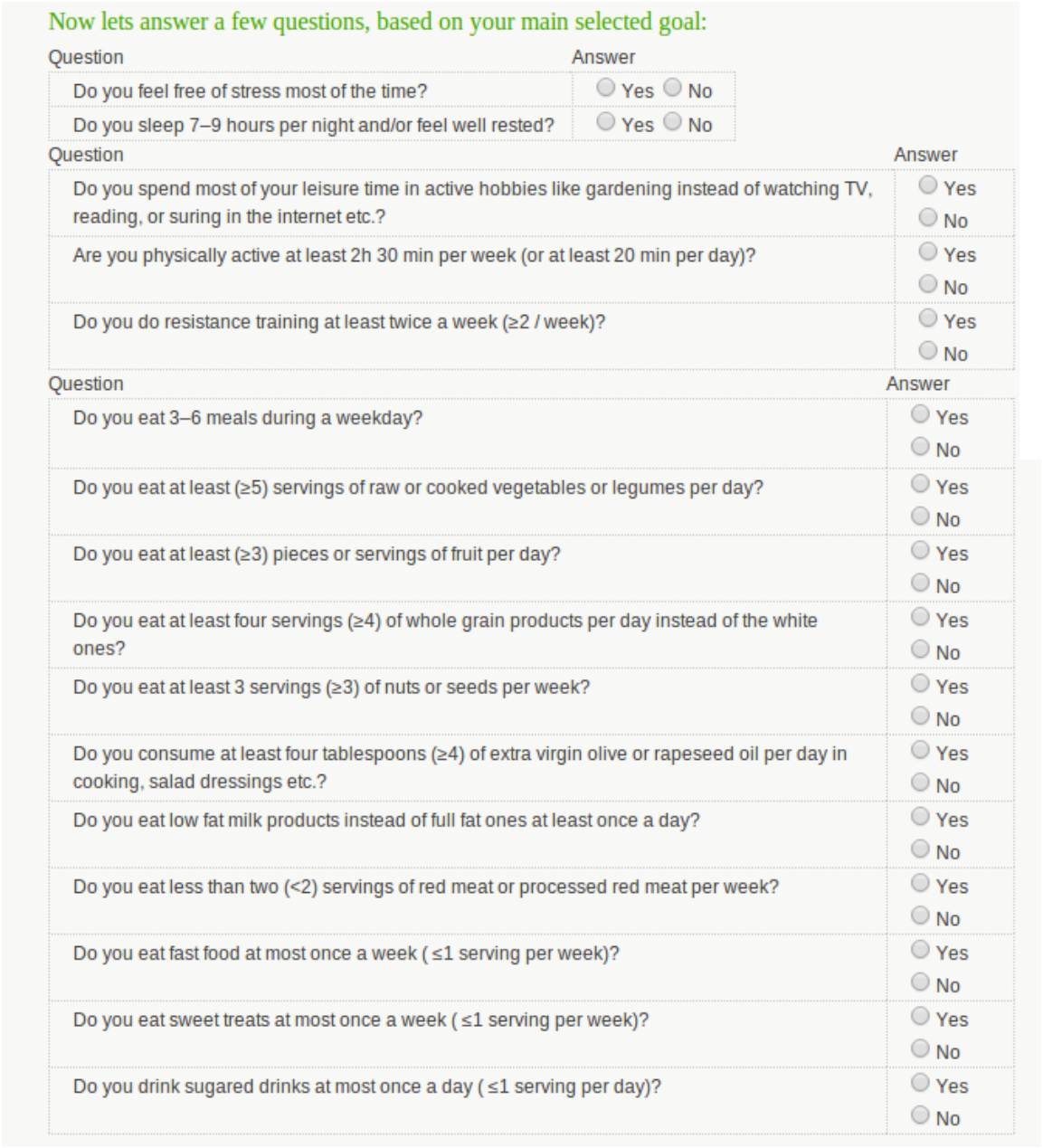
```
class Answer < ActiveRecord::Base
  attr_accessible :question_id, :text
  belongs_to :question
  has_many :translations, :as => :translatable
  accepts_nested_attributes_for :translations,:reject_if=>lambda{|a| a[:lang].blank? ||
a[:text].blank?},:allow_destroy=>true
end
```

This is a table in the database from which information is taken and the processed:

#	id	text	value	question_id	assessment_id
10	10	NULL	1	1	2
11	11	NULL	4	2	2
12	12	NULL	5	3	2
13	13	NULL	8	4	2
14	14	NULL	10	5	2
15	15	NULL	11	6	2
16	16	NULL	13	7	2
17	17	NULL	15	8	2
18	18	NULL	18	9	2
19	19	NULL	19	10	2
20	20	NULL	22	11	2
21	21	NULL	23	12	2
22	22	NULL	25	13	2

Appendix 2: Health-e-Living Questionnaire

Questionnaire that is filled out by the users when they register. After submitting users are clustered according to answers.



Now lets answer a few questions, based on your main selected goal:

Question	Answer
Do you feel free of stress most of the time?	Yes No
Do you sleep 7–9 hours per night and/or feel well rested?	Yes No

Question	Answer
Do you spend most of your leisure time in active hobbies like gardening instead of watching TV, reading, or suring in the internet etc.?	Yes No
Are you physically active at least 2h 30 min per week (or at least 20 min per day)?	Yes No
Do you do resistance training at least twice a week (≥2 / week)?	Yes No

Question	Answer
Do you eat 3–6 meals during a weekday?	Yes No
Do you eat at least (≥5) servings of raw or cooked vegetables or legumes per day?	Yes No
Do you eat at least (≥3) pieces or servings of fruit per day?	Yes No
Do you eat at least four servings (≥4) of whole grain products per day instead of the white ones?	Yes No
Do you eat at least 3 servings (≥3) of nuts or seeds per week?	Yes No
Do you consume at least four tablespoons (≥4) of extra virgin olive or rapeseed oil per day in cooking, salad dressings etc.?	Yes No
Do you eat low fat milk products instead of full fat ones at least once a day?	Yes No
Do you eat less than two (<2) servings of red meat or processed red meat per week?	Yes No
Do you eat fast food at most once a week (≤1 serving per week)?	Yes No
Do you eat sweet treats at most once a week (≤1 serving per week)?	Yes No
Do you drink sugared drinks at most once a day (≤1 serving per day)?	Yes No

Appendix 3: Helper methods

Helper methods that are used in the main controller to seperate logic and make code easier to maintain:

```
module ReportsHelper

  def calculate_points(answers, answer_points)
    points = 0
      answers.each do |answer|
          points = points + answer_points[answer] unless answer_points[answer] == nil
      end
    return points
  end

  def calculate_good_diet(answers)
       answer_points = {
                    11 => 2,
                    13 => 3,
                    15 => 5,
                    17 => 2,
                    19 => 1,
                    21 => 3}

      calculate_points(answers, answer_points)
  end
  def calculate_bad_diet(answers)
       answer_points = {
                    23 => 8,
                    25 => 3,
                    27 => 2,
                    29 => 1,
                    31 => 9}

      calculate_points(answers, answer_points)
  end
  def calculate_exercise(answers)
       answer_points = {
                    1 => 1,
                    3 => 2,
                    5 => 3,
                    7 => 7,
```

```
                9 => 1}

    calculate_points(answers, answer_points)
  end
    def calculate_membership_clusters(data_points, clusterer)
     memberships = {} # {patient_id1: clusterindex1, patient_id2: clusterindex2 for patient
2, etc }
     data_points.each do |patient_id, data_item|
      memberships[Patient.find(patient_id)] = clusterer.eval(data_item)
     end
     return memberships
    end

  def preprocess_assessments(conditions, function_x, function_y)
  data_points = {}#{patient_id1: [19,300], patient_id2: [21,20]}

    answers = Answer.find(:all, :select => 'value, assessment_id', :conditions => condi-
tions)
    grouped_answers = answers.group_by(&:assessment_id)
    grouped_answers.each do |assessment_id, answers_array|
     answer_values = answers_array.map {|a| a.value.to_i}.flatten
     xp = function_x.call(answer_values)
     yp = function_y.call(answer_values)
     assessment = Assessment.find(assessment_id)
     data_points[assessment.patient_id]=[xp,yp]
    end
    return data_points
  end

  def format_for_graph(clusters)
  cl_array=[]
  (0..clusters.length - 1).each do |i|
    cl = []
    clusters[i].data_items.sort{|a,b| a[0] <=> b[0]}.map do |a|
     cl.push({x:a[0],y:a[1]})
    end
    cl_array.push(cl)
```

```
    end
     return cl_array
    end
end
```

Appendix 4: Controller

Controller calls helper methods from another file as seen in Appendix 2:

```
include ReportsHelper
class ReportsController < ApplicationController

  respond_to :html, :js, :json

   before_filter do
   is_permitted_in?('Caregiver')
   end
   caches_action :index, :show, :layout => false

    def index
      algorithm = params[:person]
      #Retrieve, transform data, and assign to a variable
   #   @data_points = preprocess_assessments([], ReportsHelp-
er.method(params[:function_x]),
    #   ReportsHelper.method(params[:function_x]))
      @data_points = preprocess_assessments([], ReportsHelp-
er.method(:calculate_good_diet), ReportsHelper.method(:calculate_exercise))
      #Data Labels
      questions = ["x","y"]
      #Create Ai4r dataset and load data
      data_set = Ai4r::Data::DataSet.new(:data_items => @data_points.values, :data_labels
=> questions)
      #Cluster data
      clusterer = Ai4r::Clusterers::KMeans.new.build(data_set, 3)
      #Change the format of data after clustering to the format required by graph in the view
      cl_array = format_for_graph(clusterer.clusters)
      assign_memberships = calculate_membership_clusters(@data_points, clusterer)
```

```
    puts assign_memberships
    #Send data as JSON to the view
    respond_to do |format|
      format.html { render 'index' }
      format.json { respond_with cl_array}
    end
  end
end
```

Appendix 5: View

View. Contains JavaScript, styling, and some references to jQuery and Rickshaw.js libraries:

```
<script src="//ajax.googleapis.com/ajax/libs/jquery/1.7.2/jquery.min.js"
type="text/javascript"></script>
<%= stylesheet_link_tag 'rickshaw/graph.css'%>
<%= stylesheet_link_tag 'rickshaw/lines.css'%>
<%= stylesheet_link_tag 'rickshaw/detail.css'%>
<%= stylesheet_link_tag 'rickshaw/legend.css'%>
<%= javascript_include_tag 'rickshaw/rickshaw.js'%>
<%= javascript_include_tag 'rickshaw/d3.v2.js'%>
<!-- Write Javascript code here -->
<style>
#chart {
                position: relative;
                left: 40px;
                display: block;
}
#y_axis {
                position: absolute;
                top: 0;
                bottom: 0;
                width: 40px;
}
#x_axis {
                position: absolute;
```

```
                left: 40px;
                height: 40px;
}
</style>
<br/><br/><br/><br/><br/>
<select id="function_x">
  <option value="calculate_good_diet" selected="selected">Good diet</option>
  <option value="calculate_bad_diet">Bad Diet</option>
  <option value="calculate_exercise">Good Exercise</option>
</select>
<select id="function_y">
  <option value="calculate_good_diet">Good diet</option>
  <option value="calculate_bad_diet" selected="selected">Bad Diet</option>
  <option value="calculate_exercise">Good Exercise</option>
</select>
<button value="run" type="submit">Run</button>
<br />
<div id="chart_container">
                <div id="y_axis"></div>
                <div id="chart"></div>
                <div id="x_axis"></div>
                <div id="legend"></div>
</div>

<script>

var flag = 0;
$(document).ready(function(){
  $(':submit').live('click', function() { // This event fires when a button is clicked
     var button = $(this).val();
     var x_ax = $( "select#function_x").val();
     var y_ax = $( "select#function_y").val();
     if (flag = 1){
                $( "#chart" ).empty();
                $( "#x_axis" ).empty();
                $( "#y_axis" ).empty();
     };
```

```
    flag = 1;
    $.ajax({
       url: '/reports.json?function_x='+x_ax+'&function_y='+y_ax, // Action to call
       type: 'get',
       dataType: 'json',
       contentType: "application/json",
       success: function(data)
       {
                                        //alert(JSON.stringify(data));
                                        loadChart(data);
       },
       error: function (e) {
                alert("failed");
                }
    });
    return false;
 });
});

function loadChart(responseData){
// instantiate our graph!
var graph = new Rickshaw.Graph( {
            element: document.querySelector("#chart"),
            renderer: 'scatterplot',
            height: 300,
            width: 800,
            stroke: true,
            series: [
                        {
                                    data: responseData[0],
                                    color: "#c05020",
                                    name:"Cluster1"
                        }, {
                                    data: responseData[1],
                                    color: "#FCB514",
                                    name:"Cluster2"
                  }, {
```

```
                                        data: responseData[2],
                                        color: "#30c020",
                                        name:"Cluster3"
                                }
                ]
} );

var format = function(n) {

                var map = {
                                2: 'zero',
                                47: 'first'
                };

                return map[n];
}

var x_ticks = new Rickshaw.Graph.Axis.X( {
                graph: graph,
                orientation: 'bottom',
                element: document.getElementById('x_axis'),
                pixelsPerTick: 60,
                tickFormat: Rickshaw.Fixtures.Number.formatBase1024KMGTP,
} );
var y_ticks = new Rickshaw.Graph.Axis.Y( {
                graph: graph,
                orientation: 'left',
                pixelsPerTick: 60,
                tickFormat: Rickshaw.Fixtures.Number.formatBase1024KMGTP,
                element: document.getElementById('y_axis'),
} );
graph.render();
//var legend = document.querySelector('#legend');

var hoverDetail = new Rickshaw.Graph.HoverDetail( {
   graph: graph,
   xFormatter: function(x) { return x + "" },
```

```
    yFormatter: function(y) { return Math.floor(y) + " points" }
} );
};
</script>
```

Appendix 6: Test data generation

SQL script for generating test data:

```
DECLARE count INT DEFAULT 0;
 DECLARE randomnumx INT DEFAULT 0;
 DECLARE randomnumy INT DEFAULT 0;

 WHILE count < 300 DO
   /* statment */
   SET count = count + 1;
   SET randomnumx = floor(rand() * 50);
   SET randomnumy = floor(rand() * 50);
INSERT INTO test_data_table (idtest_data_table, x, y) VALUES(count, randomnumx,
randomnumy);
 END WHILE;
END
```